# Dataset Overview

|  |  |
| --- | --- |
| Field | Description |
| Dataset Name | Human Resources |
| Dataset Description | This dataset contains human resources data for a company. It includes information about employee demographics, job-related details, and satisfaction ratings. |
| Source | The source of this dataset is unknown. |
| Purpose of Use | The purpose of this dataset is to analyze factors contributing to employee attrition. By examining the relationships between various attributes and attrition, the company can gain insights into why employees leave and develop strategies to improve employee retention. |

# 2. Dataset Description

Data Structure: The dataset is structured in a tabular format, likely a CSV file or a database table. It consists of rows representing individual employees and columns representing various attributes or characteristics.

Number of Rows: 1470

Number of Columns: 35

## Variables/Features

\* `Age`: Age of the employee (Numeric - Integer)  
 \* `Attrition`: Whether the employee has left the company (Categorical - Binary: 'Yes' or 'No')  
 \* `BusinessTravel`: Frequency of business travel (Categorical: 'Travel\_Rarely', 'Travel\_Frequently', 'Non-Travel')  
 \* `DailyRate`: Daily rate of pay (Numeric - Integer)  
 \* `Department`: Department the employee belongs to (Categorical: 'Sales', 'Research & Development', 'Human Resources')  
 \* `DistanceFromHome`: Distance from home to work (Numeric - Integer)  
 \* `Education`: Education level (Numeric - Integer: 1 to 5)  
 \* `EducationField`: Field of education (Categorical: 'Life Sciences', 'Other', 'Medical', 'Marketing', 'Technical Degree', 'Human Resources')  
 \* `EmployeeCount`: Employee count (Numeric - Integer, all values are 1)  
 \* `EmployeeNumber`: Employee ID (Numeric - Integer)  
 \* `EnvironmentSatisfaction`: Employee satisfaction with the environment (Numeric - Integer: 1 to 4)  
 \* `Gender`: Gender of the employee (Categorical: 'Male' or 'Female')  
 \* `HourlyRate`: Hourly rate of pay (Numeric - Integer)  
 \* `JobInvolvement`: Job involvement level (Numeric - Integer: 1 to 4)  
 \* `JobLevel`: Job level (Numeric - Integer: 1 to 5)  
 \* `JobRole`: Job role (Categorical: 'Sales Executive', 'Research Scientist', 'Laboratory Technician', 'Manufacturing Director', 'Healthcare Representative', 'Manager', 'Sales Representative', 'Research Director', 'Human Resources')  
 \* `JobSatisfaction`: Job satisfaction level (Numeric - Integer: 1 to 4)  
 \* `MaritalStatus`: Marital status of the employee (Categorical: 'Single', 'Married', 'Divorced')  
 \* `MonthlyIncome`: Monthly income (Numeric - Integer)  
 \* `MonthlyRate`: Monthly rate (Numeric - Integer)  
 \* `NumCompaniesWorked`: Number of companies worked at (Numeric - Integer)  
 \* `Over18`: Whether the employee is over 18 (Categorical: 'Y')  
 \* `OverTime`: Whether the employee works overtime (Categorical: 'Yes' or 'No')  
 \* `PercentSalaryHike`: Percentage salary hike (Numeric - Integer)  
 \* `PerformanceRating`: Performance rating (Numeric - Integer: 3 or 4)  
 \* `RelationshipSatisfaction`: Relationship satisfaction level (Numeric - Integer: 1 to 4)  
 \* `StandardHours`: Standard hours of work (Numeric - Integer, all values are 80)  
 \* `StockOptionLevel`: Stock option level (Numeric - Integer: 0 to 3)  
 \* `TotalWorkingYears`: Total working years (Numeric - Integer)  
 \* `TrainingTimesLastYear`: Number of times employee received training in the last year (Numeric - Integer)  
 \* `WorkLifeBalance`: Work-life balance rating (Numeric - Integer: 1 to 4)  
 \* `YearsAtCompany`: Number of years at the company (Numeric - Integer)  
 \* `YearsInCurrentRole`: Number of years in the current role (Numeric - Integer)  
 \* `YearsSinceLastPromotion`: Number of years since the last promotion (Numeric - Integer)  
 \* `YearsWithCurrManager`: Number of years with the current manager (Numeric - Integer)

# 4. Potential Use Cases and Analysis

\* Descriptive Analysis:  
 \* Calculate summary statistics (mean, median, standard deviation) for numerical variables.  
 \* Analyze the distribution of categorical variables using frequency counts and percentages.  
 \* Visualize data using histograms, bar charts, pie charts, and scatter plots.  
 \* Attrition Analysis:  
 \* Identify factors that correlate with attrition (e.g., job satisfaction, work-life balance, years at the company).  
 \* Compare the characteristics of employees who have left the company with those who have stayed.  
 \* Use statistical tests (e.g., chi-square test, t-tests) to determine the significance of differences between groups.  
 \* Predictive Modeling:  
 \* Develop machine learning models to predict the likelihood of employee attrition.  
 \* Use classification algorithms such as logistic regression, decision trees, or random forests.  
 \* Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score.

# 5. Data Quality and Considerations

\* Missing Values: The dataset has no missing values.  
 \* Data Quality Issues:  
 \* The column `EmployeeCount` and `StandardHours` have the same value for all rows. These columns do not provide any meaningful information for analysis.  
 \* Potential Biases: The dataset may contain biases depending on how the data was collected and the population it represents.  
 \* Ethical Considerations:  
 \* Ensure employee privacy and data confidentiality.  
 \* Use the data responsibly and avoid discriminatory practices.

# 6. Potential Challenges

Data Imbalance: The dataset exhibits a class imbalance in the target variable (Attrition), where there are significantly more employees who have not left the company compared to those who have. This imbalance may require special techniques to address during modeling.

Complexity of Attrition: Employee attrition is influenced by a complex interplay of factors, and it may be challenging to build a highly accurate predictive model.